Research Statement Ryan Giordano

Many researchers would be concerned if they learned that some core conclusion of their statistical analysis—such as the sign or statistical significance of some key effect—could be overturned by removing a small fraction, say 0.1%, of their data. Such non-robustness would be particularly concerning if the data were not actually drawn randomly from precisely the population of interest, or if the model may have been misspecified—circumstances that often obtain in the social sciences, for example. Nevertheless, analysts do not routinely check whether ablation of such a small set could overturn their results, in part because the number of possible subsets containing 0.1% of the data points is combinatorially large.

In Broderick et al. [2020] (the author order is alphabetical; I am a lead author), I identify problematic subsets of the data using *sensitivity analysis*—that is, by forming a linear approximation to how a wide class of statistical estimators depend on their datasets. The key idea is that, although there are a very large number of subsets containing 0.1% of the data points, none of them are very different from the original dataset, and so we expect the linear approximation to work well. I confirm this intuition with finite-sample accuracy bounds in terms of intuitive and verifiable assumptions. I provide an `R` package [Giordano, 2020] to compute the approximation quickly and automatically using automatic differentiation [Baydin et al., 2017, Maclaurin et al., 2017]. And, with my co-authors, I show that the approximation is capable of detecting meaningful non-robustness in several published econometrics analyses. For example, in a study of microcredit in Mexico [Angelucci et al., 2015], we find that, by removing just 15 households out of 16,561 studied (a change of less than 0.1%), the estimated effect of microcredit changes from negative and statistically insignificant to positive and statistically significant.

In fact, my research shows that many standard, computationally demanding data analysis tasks are also amenable to fast, automatic approximation using sensitivity analysis. For example:

- Cross validation (CV) requires repeatedly leaving out subsets of the observed data and re-evaluating a statistical estimator. By forming a Taylor series approximation on the dependence of the estimator on the left-out set, I provide fast approximations to CV with finite-sample accuracy guarantees [Giordano et al., 2019b].
- Prior specification encodes key assumptions in Bayesian statistics. But Bayesian inference can be sensitive to prior specification, and evaluating the sensitivity of Bayesian posterior expectations to prior specification by re-fitting is typically computationally prohibitive due both to the large space of possible priors (often infinite dimensional), as well as the high computational cost of evaluating even a single posterior approximation. By forming a Taylor series approximation to the dependence of the posterior mean on the prior, I can explore the consequences of alternative prior functional forms at a small fraction of the cost of exact re-fitting [Giordano, 2018, Giordano et al., 2021].
- When analyzing randomly sampled data using possibly misspecified Bayesian posteriors, frequentist variability in excess of posterior variability is symptomatic of *data non-robustness*. For example, one might worry that a new random sample of poll respondents in the presidential forecast model of Gelman and Heidemanns [2020] would lead to a different prediction. This frequentist variability can be evaluated by the bootstrap, but at the considerable cost of re-running Markov Chain Monte Carlo (MCMC) hundreds of times. By approximating the dependence of the posterior on the data with sensitivity analysis, I compute accurate estimates of the frequentist variance using only a single MCMC chain—orders of magnitude faster than the bootstrap [Giordano and Broderick, 2020].
- Mean field variational Bayes (MFVB) is a popular posterior approximation method for Bayesian problems which are too large to be tractable by Markov Chain Monte Carlo [Blei et al., 2017, Regier et al., 2019]. However, MFVB approximations provide notoriously poor estimates of posterior uncertainty [Turner and Sahani, 2011]. In Giordano et al. [2018a], I show that accurate posterior covariances can be recovered from MFVB approximations with sensitivity analysis by exploiting a duality between Bayesian covariances and sensitivity.

For the remainder of this essay, I will discuss each of these applications, emphasizing the ways in which I update classical results with intuitive, relevant theory and easy-to-use computational tools.

Robustness to data ablation.

In Giordano [2020] and Broderick et al. [2020], I propose a method to assess the sensitivity of statistical analyses to the removal of a small fraction of the sample. Analyzing all possible data subsets of a certain size is computationally prohibitive, so I provide a finite-sample metric to approximately compute the number (or fraction) of observations that has the greatest influence on a given result when dropped. At minimal computational cost, our method provides an exact finite-sample lower bound on sensitivity for any estimator, so any non-robustness one finds is conclusive. I demonstrate that non-robustness to data ablation is driven by a low signal-to-noise ratio in the inference problem, is not reflected in standard errors, does not disappear asymptotically, and is not inherently a product of outliers or misspecification.

The approximation works for Z-estimators based on smooth estimating equations, a class which includes ordinary least squares, instrumental variables, generalized method of moments, variational Bayes, and maximum likelihood estimators. Using my `R` package [Giordano, 2020], the approximation is automatically computable from the specification of the estimating equation alone. By analyzing several published econometric analyses [Angelucci and De Giorgi, 2009, Finkelstein et al., 2012, Meager, 2019], I show that even 2-parameter linear regression analyses of randomized trials can be highly sensitive. While I find some applications are robust, in others the sign of a treatment effect can be changed by dropping less than 1% of the sample even when standard errors are small.

Approximate cross validation.

The error or variability of machine learning algorithms is often assessed by repeatedly re-fitting a model with different weighted versions of the observed data; cross-validation (CV) can be thought of as a particularly popular example of this technique. In Giordano et al. [2019b], I use a linear approximation to the dependence of the fitting procedure on the weights, producing results that can be faster by an order of magnitude than repeated re-fitting. I provide explicit finite-sample error bounds for the approximation in terms of a small number of simple, verifiable assumptions. My results apply whether the weights and data are stochastic or deterministic, and so can be used as a tool for proving the accuracy of the infinitesimal jackknife on a wide variety of problems. As a corollary, I state mild regularity conditions under which the approximation consistently estimates true leave-k-out cross-validation for any fixed k. I demonstrate the accuracy of the approximation on a range of simulated and real datasets, including an unsupervised clustering problem from genomics [Luan and Li, 2003, Shoemaker et al., 2015].

Approximately bootstrapping Bayesian posterior means.

The frequentist (i.e., sampling) variance of Bayesian posterior expectations differs in general from the posterior variance even for large datasets, particularly when the model is misspecified or contains many latent variables [Kleijn and van der Vaart, 2006]. Unlike the posterior variance, the frequentist variance is meaningful even in the presence of misspecification, particularly when the data is known to arise from random sampling [Waddell et al., 2002]. However, the principal existing approach for computing the frequentist variability of MCMC procedures is the bootstrap, which can be extremely computationally intensive due to the need to run hundreds of extra MCMC procedures [Huggins and Miller, 2019].

In Giordano and Broderick [2021, 2020], I propose an efficient alternative to bootstrapping an MCMC procedure. My approach is based on the Bayesian analogue of the influence function from the classical frequentist robustness literature. Using results from [Giordano et al., 2018a, 2019b], I show that the influence function for posterior expectations can be easily computed from the posterior samples of a single MCMC procedure and consistently estimates the bootstrap variance. I demonstrate the accuracy and computational benefits of the influence function variance estimates on array of experiments including an election forecasting model [Gelman and Heidemanns, 2020], the Cormack-Jolly-Seber model from ecology [Kéry and Schaub, 2011], and a large collection of models and datasets from the social sciences [Gelman and Hill, 2006].

Bayesian sensitivity analysis.

Prior sensitivity for Markov Chain Monte Carlo. MCMC is arguably the most commonly used computational tool to estimate Bayesian posteriors, which is made still easier by modern black-box MCMC tools such as `Stan` [Carpenter et al., 2017, Stan Development Team, 2020]. However, a single run of MCMC typically remains time-consuming, and systematically exploring alternative prior parameterizations by re-running MCMC would be computationally prohibitive for all but the simplest models.

My software package, `rstansensitivity`, [Giordano, 2018, Giordano et al., 2018b], takes advantage of the automatic differentiation capacities of `Stan` [Carpenter et al., 2015] together with a classical result from Bayesian robustness [Gustafson, 1996, Basu et al., 1996, Giordano et al., 2018a] to provide automatic hyperparameter sensitivity for generic `Stan` models from only a single MCMC run. I demonstrate the speed and utility of the package in detecting excess prior sensitivity, particularly in a social sciences model taken from Gelman and Hill [2006, Chapter 13.5].

Prior sensitivity for discrete Bayesian nonparametrics. A central question in many probabilistic clustering problems is how many distinct clusters are present in a particular dataset and which observations cluster together. Discrete Bayesian nonparametric (BNP) mixture models address this question by placing a generative process on cluster assignment, making the number of distinct clusters present amenable to Bayesian inference. However, like all Bayesian approaches, BNP requires the specification of a prior, and this prior may favor a greater or lesser number of distinct clusters.

In Giordano et al. [2021], I derive and analyze prior sensitivity measures for variational Bayes approximations in general, with a focus on discrete BNP models. Unlike much previous work on local Bayesian sensitivity for BNP (e.g. Basu [2000]), I pay special attention to the ability of the sensitivity measures to *extrapolate* to different priors, rather than treating the sensitivity as a measure of robustness *per se*. Further, though there are many valid choices for measuring the "size" of a prior perturbation for exact Bayesian posteriors [Gustafson, 1996], I prove that, for variational Bayes approximations, local approximations are valid only when based on the largest point-wise change in the log density. My co-authors and I apply our sensitivity measures to a number of real-world problems, including an unsupervised clustering problem from genomics using fastSTRUCTURE [Raj et al., 2014], demonstrating that the approximation is accurate, orders of magnitude faster than re-fitting, and capable of detecting meaningful prior sensitivity.

Uncertainty propagation in mean-field variational Bayes.

Mean-field Variational Bayes (MFVB) is an approximate Bayesian posterior inference technique that is increasingly popular due to its fast runtimes on large-scale scientific data sets (e.g., Raj et al. [2014], Kucukelbir et al. [2017], Regier et al. [2019]). However, even when MFVB provides accurate posterior means for certain parameters, it often mis-estimates variances and covariances [Wang and Titterington, 2004, Turner and Sahani, 2011] due to its inability to propagate Bayesian uncertainty between statistical parameters.

In Giordano et al. [2015, 2018a], I derive a simple formula for the effect of infinitesimal perturbations on MFVB posterior means, thus providing improved covariance estimates and greatly expanding the practical usefulness of MFVB posterior approximations. The estimates for MFVB posterior covariances rely on a result from the classical Bayesian robustness literature that relates derivatives of posterior expectations to posterior covariances and includes the Laplace approximation as a special case. In the experiments, I demonstrate that my methods are simple, general, and fast, providing accurate posterior uncertainty estimates and robustness measures with runtimes that can be an order of magnitude faster than MCMC, including models from ecology [Kéry and Schaub, 2011], the social sciences [Gelman and Hill, 2006], and on a massive internet advertising dataset [Criteo Labs, 2014].

Selected Future work

My research is ideally driven by the needs of my scientific and industry collaborators, and so I expect my future work will be determined to a large part by my colleagues, and so inherently difficult to predict in advance. Additionally, many of my existing papers can be easily "crossed" with one another, producing relatively straightforward projects suitable for more junior researchers. For example, by combining the Bayesian influence function of Giordano and Broderick [2021] with the adversarial data ablation metrics of Broderick et al. [2020], one can automatically find influential data subsets from MCMC output. (I am presently pursuing this idea with the `R` package `rstanarm` together with a PhD student.)

That aside, there are a few interesting directions that I find promising and interesting.

The empirical influence function (EIF). Much of my work (particularly Giordano et al. [2019b], Broderick et al. [2020], Giordano and Broderick [2021]) has strong connections to the classical theory of von Mises expansions and the closely related concept of the "influence function," which measures the effect of individual datapoints on an estimator [Mises, 1947, Reeds, 1976, Hampel, 1986, Serfling, 2009]. But my focus on the the influence function evaluated at the observed data—i.e., the "empirical influence function" (EIF)—stands in contrast with much of the classical literature, which studies asymptotic behavior via the (unobserved) limiting influence function .

In our present age of automatic differentiation, large datasets, and complex models, I believe that the EIF will continue to provide practical benefits and is relatively under-studied. Since the EIF is observed, one can in principal use it to compute exact finite-sample error bounds for its approximations. [1] Studying the EIF alone avoids many of the technical difficulties requried for classical von Mises calculus, since the latter must study distribution functions in spaces that embed both continuous and discrete measures [Gill et al., 1989].

In Giordano et al. [2019a], I show that higher-order EIFs can be easily and automatically evaluated and analyzed for M-estimators at a computational cost comparable to the first-order EIF, a result which I find particularly exciting. I will describe three tentative but promising directions. For example, one can use a higher-order EIF to efficiently approximate the bootstrap more accurately than the bootstrap approximates the true sampling distribution, opening up the possibility of tractable approximations to bootstrap-after-bootstrap procedures that are currently computatoinally prohibitive [Shao and Tu, 2012]. Additionally, I believe that one can use higher-order EIFs to study the sampling uncertainty of cross-validation: as I show in Giordano et al. [2019b], cross-validation can be well-approximated with a first-order EIF. It follows that the sampling uncertainty in cross-validation could be estimated via a von-Mises expansion of the first-order EIF, which is itself a second-order expansion. It may be that higher-order EIFs open the door to computable approximations to higher-order jackknife bias corrections, which can be valuable but which are combinatorially difficult to compute in all but special circumstances [Burnham and Overton, 1978].

Sensitivity analysis in difficult situations. It is not always as easy to apply sensitivity analysis in practice as it is in theory. I have found that a few key problems tend to recur, and I will discuss them in turn, as well as potential solutions which draw connections to the optimization literature.

- Non-smooth objectives (e.g., the lasso penalty),
- Incomplete optimization, and / or
- Large and / or difficult to compute matrices of second derivatives.

For non-smooth objectives, a promising idea is to use local approximations to speed up computationally intensive but smooth components in non-smooth problems. Consider, for example, cross-validating linear regression with a lasso penalty: the optimal squared error is a smooth function and the lasso penalty is non-smooth. As observed by Wilson et al. [2020], one can then form a fast approximation to

[1] Though I provide finite-sample error bounds in Giordano et al. [2019b,a], the bounds need to be tightened and simplified to be practically useful. There is good reason to believe they can be, especially in simple cases such as least-squares estimators.

the cross-validated loss by forming a fast approximation to the effect on the optimal squared error of leaving out a single datapoint, and retaining non-smoothness in the lasso penalty intact.

A collaborator from biostatistics is currently working with me to apply my work in Broderick et al. [2020] to single-cell sequencing data in hopes of flagging small groups of highly influential cells. Careful analysis of the output of the `R` package `DESeq2` [Love et al., 2014] shows that the package's iteratively reweighted least squares algorithm may not be finding a high-quality optimum of the objective function. The "sensitivity" of incomplete optimization is not well-defined, but we ideally do not want to require end-users to to re-implement their optimization algorithm to use sensitivity analysis. One potential solution using higher-order expansions could be to represent incomplete optimization as a constrained optimization problem in which the gradient is set not to zero, but to the value numerically observed in the incomplete optimization. Representing the constraint with a Lagrange multiplier, one can then use a second-order expansion to estimate the effect of simultaneously removing the constraint (setting the multiplier to zero) and perturbing the objective (e.g., removing some data).

Finally, the key computational bottleneck in sensitivity analysis for M-estimators is the solution of linear systems involving the inverse Hessian of the objective function. Indeed, as I show in Giordano et al. [2019a], influence functions of all orders share essentially the same bottleneck. When the Hessian matrix is too large to store in memory (much less factorize), I have gotten good results using off-the-shelf iterative algorithms like the conjugate-gradient algorithm [Nocedal and Wright, 2006], but there is surely more work to be done speeding up this step of the procedure. Stochastic second-order methods are currently an active research topic in optimization [Agarwal et al., 2017, Berahas et al., 2020], and methods developed therein should apply directly to sensitivity analysis.

References

Agarwal, N., Bullins, B., and Hazan, E. (2017). Second-order stochastic optimization for machine learning in linear time. *The Journal of Machine Learning Research*, 18(1):4148–4187.

Angelucci, M. and De Giorgi, G. (2009). Indirect effects of an aid program: How do cash transfers affect ineligibles' consumption? *American Economic Review*, 99(1):486–508.

Angelucci, M., Karlan, D., and Zinman, J. (2015). Microcredit impacts: Evidence from a randomized microcredit program placement experiment by Compartamos Banco. *American Economic Journal: Applied Economics*, 7(1):151–82.

Basu, S. (2000). Bayesian robustness and Bayesian nonparametrics. In Insua, D. R. and Ruggeri, F., editors, *Robust Bayesian Analysis*, volume 152. Springer Science & Business Media.

Basu, S., Jammalamadaka, S. R., and Liu, W. (1996). Local posterior robustness with parametric priors: Maximum and average sensitivity. In *Maximum Entropy and Bayesian Methods*, pages 97–106. Springer.

Baydin, A., Pearlmutter, B., Radul, A., and Siskind, J. (2017). Automatic differentiation in machine learning: A survey. *Journal of Machine Learning Research*, 18(153):1–153.

Berahas, A., Bollapragada, R., and Nocedal, J. (2020). An investigation of Newton-sketch and subsampled Newton methods. *Optimization Methods and Software*, pages 1–20.

Blei, D., Kucukelbir, A., and McAuliffe, J. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Broderick, T., Giordano, R., and Meager, R. (2020). An automatic finite-sample robustness metric: Can dropping a little data change conclusions? *arXiv preprint arXiv:2011.14999*. Note: Following conventions in econometrics, the authors are listed alphabetically. Giordano and Meager are equal contribution primary authors.

Burnham, K. and Overton, W. (1978). Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika*, 65(3):625–633.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, 76(1).

Carpenter, B., Hoffman, M., Brubaker, M., Lee, D., Li, P., and Betancourt, M. (2015). The Stan math library: Reverse-mode automatic differentiation in C++. *arXiv preprint arXiv:1509.07164*.

Criteo Labs (2014). Criteo conversion logs dataset. Downloaded on July 27th, 2017.

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J., Allen, H., Baicker, K., and Oregon Health Study Group (2012). The Oregon health insurance experiment: Evidence from the first year. *The Quarterly Journal of Economics*, 127(3):1057–1106.

Gelman, A. and Heidemanns, M. (2020). The Economist: Forecasting the US elections. Data and model accessed Oct., 2020.

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gill, R., Wellner, J., and Præstgaard, J. (1989). Non-and semi-parametric maximum likelihood estimators and the von mises method (part 1). *Scandinavian journal of statistics*, pages 97–128.

Giordano, R. (2018). StanSensitivity: Automated hyperparameter sensitivity for Stan models. GitHub repository `https://github.com/rgiordan/StanSensitivity`.

Giordano, R. (2020). Zaminfluence. GitHub repository `https://github.com/rgiordan/zaminfluence`.

Giordano, R. and Broderick, T. (2020). Effortless frequentist covariances of posterior expectations in Stan. Presentation at Stancon 2020 `https://tinyurl.com/y2e2ucp3`.

Giordano, R. and Broderick, T. (2021). The Bayesian infinitesimal jackknife for variance. *In preparation.*

Giordano, R., Broderick, T., and Jordan, M. (2018a). Covariances, robustness and variational Bayes. *The Journal of Machine Learning Research*, 19(1):1981–2029.

Giordano, R., Broderick, T., and Jordan, M. I. (2015). Linear response methods for accurate covariance estimates from mean field variational Bayes. In *Advances in Neural Information Processing Systems*, pages 1441–1449.

Giordano, R., Broderick, T., and Jordan, M. I. (2018b). Automatic robustness measures in Stan. Presentation at Stancon 2018 `https://tinyurl.com/yyqwpowc`.

Giordano, R., Jordan, M. I., and Broderick, T. (2019a). A higher-order Swiss army infinitesimal jackknife. *arXiv preprint arXiv:1907.12116*.

Giordano, R., Liu, R., Jordan, M. I., and Broderick, T. (2021). Evaluating sensitivity to the stick breaking prior in Bayesian nonparametrics. *https://arxiv.org/abs/2107.03584*.

Giordano, R., Stephenson, W., Liu, R., Jordan, M. I., and Broderick, T. (2019b). A Swiss army infinitesimal jackknife. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1139–1147.

Gustafson, P. (1996). Local sensitivity of posterior expectations. *The Annals of Statistics*, 24(1):174–195.

Hampel, F. (1986). *Robust statistics: The approach based on influence functions*, volume 196. Wiley-Interscience.

Huggins, J. and Miller, J. (2019). Using bagged posteriors for robust inference and model criticism. *arXiv preprint arXiv:1912.07104*.

Kéry, M. and Schaub, M. (2011). *Bayesian population analysis using WinBUGS: A hierarchical perspective*. Academic Press.

Kleijn, B. and van der Vaart, A. (2006). Misspecification in infinite-dimensional Bayesian statistics. *The Annals of Statistics*, 34(2):837–877.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. (2017). Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474.

Love, M., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15:550.

Luan, Y. and Li, H. (2003). Clustering of time-course gene expression data using a mixed-effects model with B-splines. *Bioinformatics*, 19(4):474–482.

Maclaurin, D., Duvenaud, D., and Johnson, M. (2017). autograd. GitHub repository `https://github.com/HIPS/autograd`.

Meager, R. (2019). Understanding the average impact of microcredit expansions: A Bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics*, 11(1):57–91.

Mises, R. (1947). On the asymptotic distribution of differentiable statistical functions. *The Annals of Mathematical Statistics*, 18(3):309–348.

Nocedal, J. and Wright, S. (2006). *Numerical optimization.* Springer Science & Business Media.

Raj, A., Stephens, M., and Pritchard, J. (2014). fastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics*, 197(2):573–589.

Reeds, J. (1976). *On the definition of von Mises functionals.* PhD thesis, Ph. D. Thesis, Statistics, Harvard University.

Regier, J., Fischer, K., Pamnany, K., Noack, A., Revels, J., Lam, M., Howard, S., Giordano, R., Schlegel, D., and McAuliffe, J. (2019). Cataloging the visible universe through Bayesian inference in Julia at petascale. *Journal of Parallel and Distributed Computing*, 127:89–104.

Serfling, R. (2009). *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons.

Shao, J. and Tu, D. (2012). *The Jackknife and Bootstrap.* Springer Series in Statistics.

Shoemaker, J. E., Fukuyama, S., Eisfeld, A. J., Zhao, D., Kawakami, E., Sakabe, S., Maemura, T., Gorai, T., Katsura, H., Muramoto, Y., Watanabe, S., Watanabe, T., Fuji, K., Matsuoka, Y., Kitano, H., and Kawaoka, Y. (2015). An ultrasensitive mechanism regulates influenza virus-induced inflammation. *PLoS Pathogens*, 11(6):1–25.

Stan Development Team (2020). RStan: the R interface to Stan. R package version 2.21.2.

Turner, R. E. and Sahani, M. (2011). Two problems with variational expectation maximisation for time-series models. In Barber, D., Cemgil, A. T., and Chiappa, S., editors, *Bayesian Time Series Models.*

Waddell, P., Kishino, H., and Ota, R. (2002). Very fast algorithms for evaluating the stability of ML and Bayesian phylogenetic trees from sequence data. *Genome Informatics*, 13:82–92.

Wang, B. and Titterington, M. (2004). Inadequacy of interval estimates corresponding to variational Bayesian approximations. In *Workshop on Artificial Intelligence and Statistics*, pages 373–380.

Wilson, A., Kasy, M., and Mackey, L. (2020). Approximate cross-validation: Guarantees for model assessment and selection. In *International Conference on Artificial Intelligence and Statistics*, pages 4530–4540. PMLR.